# AN EDITOR'S COMMENTS ON 'REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM', WHITE BOOK VERSION 2.0

BY DMS 97-10-18

These comments begin by giving a view on the major influences to the changes from White Book 1 to White Book 2, and the general status of White Book 2. (Note that there is a Word 6 version of White Book 2 that includes the revision marks as compared with White Book 1.) Subsequently we provide a list of the Silver Spring agreements on updates, organized by document section, and the extent to which they are reflected in White Book - 2. Finally, we summarize a number of known issues.

## 1 Major Influencing Items:

A major item influencing the changes between WB-1 and WB-2 has been the Silver Spring (SP) recommendations on making section 2 into an overview of the major concepts, with reduced and re-organized material that does not use OMT diagrams. We feel this has worked very well, but it has also significantly impacted the organization of the section 4.2 information model material (previously section 3.2).

Another major item has been the impacts of incorporating our SP decision to make clearer that all the information objects have Representation Information, together with the decision to call the lowest level Information Package an Information Unit. We did this by keeping the simple view of SIPs going to AIPs going to DIPs, but allowing the AIPs to be specialized in one of two forms - either an AIU (lowest level package going to/from Archival Storage) or an AIC (collection of AIUs and/or other AICs). Some at the recent US workshop found this confusing and some found it useful. This needs discussion.

A third major item has been our effort to clarify the migration views as they impact Archival Storage and as they relate to SIP to AIP and AIP to DIP conversions. This resulted in our recognizing that we needed to be able to talk about another category of information, in relation to the Information Package, that is used to bind and identify the components of such packages. We called this Packaging Information and it has a major role to play. It is needed to find the SIP and DIP components in any transmissions, and it is modified in some types of migrations within the archive.

The collection of information modeling terminology changes has impacted much of the document and necessitated updates, in particular, to the text in the detailed functional model of section 4.1 (old section 3.1). This material was not well integrated in the use of the data model terms in WB 1 to begin with, so much of section 4.1 has new detailed text even though the same concepts are being expressed. This brings up an issue in that it now looks like the document would flow more smoothly if the data model discussion followed a brief introduction to the functional entities and preceded the detailed functional model discussion, much as has been done in section 2. This would allow an even cleaner use of terms in the detailed functional model section.

Finally, there has long been Panel and US concern with the inadequacy of the material addressing Customer access to the archive's information. Following the last US

workshop, as Don and Lou were attempting to pull all of this together, a better way to represent this information from a modeling perspective was recognized. We find it is simpler and more elegant, and it has allowed us to discuss the issues much more clearly and completely. It has been incorporated and explained with the use of an archive of digital movies. However, since this material is very new, it has not been tested and may very well need clarifications. It has resulted in some changed terms, such as Collection Descriptors and Unit Descriptors, that have not been fully reflected in the detailed functional model section (nor could they be unless the information model moves ahead of this section). Some of the related terms, which are new, may need improvement.

## 2 Agreed Changes and Their Status

These changes come from the Silver Spring minutes (SPM), from the agreed items in the Silver Spring paper N12 on US proposed edits (USE), and from the agreed responses to the CNES comments (USC) in paper N10 and BNSC comments (USB) in paper N11.

There have been several contributors to updated sections. Don has provided updates to sections 1, 2 and 3. Mike Martin has been the primary editor for section 4.1 functional descriptions and Paul Grunberger has provided the section 4.1.8 data flow diagrams. Lou has provided updates to sections 4.2 and 4.3 on information models. Randy Davis has provided updates to the migration section and the illustrative scenario. Don added substantial information model terminology to Randy's migration material at the last minute. There have been no substantial improvements to the Classification material, which is still weak. Bruce Ambacher has updated the NARA scenario and provided another scenario on the PERINATAL project. Alan Wood has provided an updated Life Sciences scenario, and Claude Huc has provided a scenario on the developing CNES data center. Paul Grunberger and Mike Martin have cooperated on generating the Matrix in Annex C. Paul has provided the federation material (initial draft) in Annex D. Don provided the material on OMT based on a cut that ISO TC211 did on an OMT aid. Don and Lou completed the editing of version 2.

A few minor edits have not been addressed due to an inadequate editing resource. We expect this problem to be behind us with the new fiscal year.

### 2.1 Boiler Plate Material

Item 1: It was agreed that the White Book should include an abstract in the forward (SPM)
Status: This is inappropriate under the CCSDS and ISO style guides and has not been incorporated.

### 2.2 Purpose and Scope (section 1.1)

Item 1: In second bullet item, change "architecture's" to "architectures"(SPM)
Status: Completed - now bullet 3.

Item 2: In second bullet, break out a separate bullet that says: "- provide a basis for comparing the data modeling and transformation aspects of digital information preserved and managed by these archives;" (SPM)
Status: Completed, but revised to be clearer as : "provides a basis for comparing the data models of digital information preserved by archives and for discussing how data models and the underlying information may change over time."

Item 3: clarify that digital information includes digital Metadata (SPM)
Status: Completed, without using term 'metadata'

Item 4: clarify that OAIS is also applicable to non-digital information (e.g. physical samples) since non-digital information has digital Metadata (SPM)
Status: Completed

Item 5: address exploitation of the information preserved (access, dissemination)(SPM)
Status: Completed

Item 6: change 4th bullet into change consensus on the elements and processes(SPM)
Status: Completed - now the 6th bullet

Item 7: change 5th bullet to guide the identification and production(SPM)
Status: Completed - now the 7th bullet

Item 8: rephrase minimum set of responsibilities and first step (SPM)
Status: Completed

Item 9: in second bullet, break out a separate bullet according to US Workshop proposal(SPM)
Status: Completed

Item 10: the first use of the term "long term" in purpose and scope should be longer defined (SPM)
Status: It has been more completely defined.

### 2.3 Applicability (section 1.2)

Item 1: provide clear distinction organisation that might be temporary, but might be responsible for long term preservation (SPM)
Status: Completed

Item 2: use only the term 'long term' throughout the whole document (SPM)
Status: This is used primarily, but there may be a few places where 'indefinite long term' is still used.

Item 3: the words permanent and indefinite will be used in the definition of long term in glossary (SPM)
Status: " Indefinite" is used

### 2.4 Rationale (section 1.3)

Item 1: model section 1.3 after the information content in 2.1 first two paragraphs (SPM)
Status: editor (DS) finds that this does not work as it makes the material very redundant. Editor recommends leaving it as is.

### 2.5 Road Map (section 1.4)

Item 1: shorten the paragraphs in 1.4 (5th bullet) by deleting the last two sentences (SPM)
Status: Completed

Item 2: address data exchange, metadata exchange and cross archive browsing (distributed archives) when interoperability is discussed in the document (SPM)
Status: An annex on Federated systems has been generated. It is new, has not been reviewed, but an updated version could become a new section of the document.

Item 3: add new bullet to 1.4: standards for digital data identification in the archive (SPM)
Status: Completed

Item 4: modify glossary so that user is not only human, but also machine (SPM)
Status: Client has been added to the glossary, and a Consumer or Producer can be either a person or a Client.

Item 5: delete in 1.4 bullet 3 "to data users" (SPM)
Status: Completed

Item 6: add an expanded version of section 1.4 to the document (SPM)
Status: Not completed due to a lack of proposed material. What else is desired?

Item 7: add example to 7th bullet "specific media" (SPM)
Status: OPEN - however an example might be 'optical tape error correction reporting', or 'a description language for the format of data laid out on digital linear tape by mass storage systems'.

### 2.6 Definitions (section 1.6)

Item 1: change definition of consumer so that consumer can be also a kind of client (SPM)
Status: Completed

Item 2: add client to list of terms. That will also facilitate the description of finding aids.(SPM)
Status: Completed; Finding Aid should be updated to replace 'user' with 'Consumer'

Item 3: add designated community to definitions (SPM)
Status: Completed

Note: There has been substantial additional updates for consistency

### 2.7 OAIS Concepts (section 2)

Item 1: section 2 is a concept description and section 3 defines the "Model" (SPM)
Status: This tack has been taken, where section 2 attempts to give the major concepts and section 4 (was 3) gives a more detailed view.

Item 2: page 2-4 and 2-5 needs to be moved (SPM)
Status: material moved to detailed information model (4.2)

Item: 3 delete Figure 2-2 and 2-3 (SPM)
Status: Completed

Item 4: remove all OMT in section 2. (SPM)
Status: Completed

Item 5: reduce section 2.2.1 to size of section 2.2.2 (SPM)
Status: Completed - one page on 'information', one on 'information package' and one on 'information package variants'

Item 6: remove example in section 2.2.1 (SPM)

Status: Extended example and representation information discussion are removed. However two short examples are included in the one page 'information definition'

Item 7: DS will write the key ideas on representation information and include them in next White Book Version (SPM)
Status: Section 2 defines information, and an information object, which has representation information. Two short examples, one physical and one digital, are given. An expanded discussion on representation, combining much of the previous section 2 and 4 materials, and is given in Section 4.2.

Item 8: Examples on representation information to be included in the next version of white book (see DS/9705/P2/8/ and GMP/9705/P2/9) (SPM)
Status: Examples are given, but are they adequate?

Item 9: In Section 2. a brief overview should be included how user access the data, by July 1. (SPM)
Status: This sections previously had this material, and it has been updated slightly. Is there something more desired?

Item 10**:** Move material in section 2.3 before 2.2, by July 1. (SPM)
Status: Completed, for the introduction of the OAIS environment. This is followed by the Information Model views, which is followed by a more detailed discussion of the OAIS environment that uses the information package terms with a figure. Note: there is a proposal to take 'Result Sets' out of the definitions and replace it in the figure with simply 'query results'.

Item 11: Update paragraph under Figure 2-5 (on page 2-8) and first paragraph on page 2-9 by July 1 (SPM)
Status: The discussion under the Environmental figure (now figure 2-1) has been clarified and expanded.

Item 12: LR provide more details in next White Book on levels of interoperability amongst archives and federated archives concept, by September 30 (SPM)
Status: Paul Grunberger (US support) has provided initial material and it is found in Annex B.

Item 13: Change the text in 2.3.2 to reflect more that they are activities than classes. Make bullets from text, by July 1 (SPM)
Status: Completed - still in section 2.3.2

Item 14: change 'preservation description information' to 'preservation information', so to be consistent with all other terms**.** (USB)
Status: We started to do this with version 1.1, but stopped for a couple of reasons:
- the acronym is PI, which is now the same as Packaging Information (PI)
- preservation information may be confused with 'preserved information'; it would be more correct to call it 'preserving information', but this still has the PI acronym.

Item 15: Section 2.2.2 is an important section of the document and I believe more detailed examples for the different categories of 'preservation information' is needed**.** (USB)
Status: OPEN - More detailed examples of the Preservation Description Information have not yet been added.

### 2.8 OAIS Responsibilities (section 3)

Item 1: section 2.4 bullet list remains in section 2 and make rest of 2.4 a complete new section that addresses CN's and DG's concerns. (SPM)
Status: Editor does not find that this works well. The remaining material is a direct and parallel expansion of the bullets and therefore it does not work to split this material. The US WS recommended that the bullets also be included in the new Section 3, and this has been completed.

Item 2: Add to 2.4.2 sentence according to the following lines: "determining the designated community needs to include evolution perspectives" (SPM)
Status: Completed, in what is now section 3.2.

Item 3: Add to 2.4.6 some sentences about access control, by July 1. (SPM)
Status: Completed, in what is now section 3.6.

### 2.9 Functional Model (section 4.1)

General comment and issue: The text in this section has been rewritten to fit within a reduced number of sub-functions, and to better reflect the information model terminology. The items mentioned in the text that flow between entities are indicated in bold, with a following identifier in brackets, to assist in linking them to the data flows of figures 4-2 and 4-3. This technique has been a US discussion item as to whether this should be retained in the final document, and if so, how the identifiers should be constructed.

Item 1: **r**emove in Figure 3-1 the dotted lines and explain that the lines in the figure refer to an interface between the two functions. (SPM)
Status: Completed - now figure 4-1.

Item 2: change in section 3.1.2. 'data preparer' into 'data producer' (SPM)
Status: Completed - now section 4.1.2

Item 3: in 3.1.2 first paragraph replace 'other' in the statement 'reviewed by the archive staff' and substitute 'other' by 'automated tools and others' (SPM)
Status: OPEN - The 'others' was meant to refer to 'possibly external reviewers from the Designated Community'. It can include automated tools.

Item 4: ingest does validation/verification and administration does review (SPM)
Status: OPEN - Editor and US workshop feels it is more natural to think of all of this activity as a part of Ingest, which is very extensive in many cases and requires discipline experts to participate in some conversions, for example. See the current write-up on the 'review' subfunction. This is not to suggest that Administration doesn't have the final say, but we think this is not the main activity of interest here.

Item 5: change title of section 3.1.3 to 'Storage' only (SPM)
Status: This was to make the section consistent with Figure 4-1. However, our experience is that others confuse 'storage' with such things as 'cashing storage, etc., and therefore we have gone back to 'Archival Storage' to differentiate it from general storage.

Item 6: transfer (page 3-3) change third sentence into 'this might be either an electronic, logical or physical transfer'. Staging area is for logical transfer s in the same space (metadata change). (SPM)
Status: Currently this reads, on page 31 under Transfer Initiation: "This may be either an electronic, physical, or virtual (i.e., the data stays in one place) transfer." We believe the staging area is for electronic, and possibly for physical, transfers. We could certainly substitute 'logical' for 'virtual'.

Item 7: two typos in 3.1.2. conversion : ‘Conversions’ and ‘motored’ (SPM)
Status: Corrected - section 4.1.2

Item 8: delete ‘and data objects that comprise them’ in first sentence of 3.1.3 (SPM)
Status: Corrected - 4.1.3

Item 9: ‘error checking’ (p 3-4) should include also a checking function for degeneration in the storage. (SPM)
Status: OPEN - Need discussion on what is seen as practical in this regard.(section 4.1.3)

Item 10: include some examples for Report Request in 3.1.4 (SPM) and Report Generation (USC)
Status: OPEN - this was missed in section 4.1.4. However, a Report Request is just a request to get any information from the Data Management persistent storage (data base) including all Descriptors, statistics, Customer profiles, etc. It is a query, although it may be handled by persons as well as by automated systems. The Report Generation is the delivery of this information.

Item 11: in 3.1.5 use same style format as in rest of document (SPM)
Status: Styles are made consistent throughout section 4.1

Item 12: change 3.1.5. all requirements (‘shall’) into paragraphs (SPM)
Status: The ‘shalls’ have been removed in section 4.1.

Item 13: include access control planning (SPM)
Status: OPEN - this is not currently indicated explicitly. This could be added to Planning and Scheduling of Administration.

Item 14: in administration more is needed for user management: registration, subscription, statistical data of users, special privileges, creation of user passwords/names, etc. (SPM)
Status: See the Customer Service description of Access.

Item 15: add in 3.1.5 text about ‘submission agreement’ (SPM)
Status: OPEN - Currently this is in Ingest. Should this be primarily an Administration function, or should it be in Ingest where there will be the expertise needed to handle most Submission Agreements, apart from formal approvals?

Item 16: advanced development in 3.1.6 should be deleted in Access and moved to ‘configuration control’ in 3.1.5 (SPM)
Status: There is a Prepare Finding Aids function in Access because, again, we think it is more natural to associate Access with this function than with Administration.

Item 17: suppress Boolean in ‘Query’ in 3.1.6 (SPM)
Status: Completed - section 4.1.6.

Item 18: access should have cost estimation (SPM)
Status: OPEN - This was lost and should be made a service visible to the Customer.

Item 19: in section 3. needs to elaborate access control in 3.1.6 (SPM)
Status: OPEN - Desire was to include concepts of proprietary access and volume control. These are not mentioned explicitly. Do they need to be?

Item 20: make 3.1.7 ‘confirm delivery ‘ more general. (SPM)
Status: Completed - 4.1.7

Item 21: re-write ‘off-line’ and ‘on-line’, since not really adequate terms. If session is maintained than this is on-line. (SPM)
Status: Completed - 4.1.4

Item 22: move 3.1.8. into annex and provide more text, by July 1 (SPM)
Status: Matrix has been moved into Annex C. No additional text has been provided, but it does reflect the reduction in the number of categories and there is a more meaningful alignment.

Item 23: The diagrams in 3.1.9. will be substituted by the SIL/97/P2/N9 (SPM)
Status: The data flow and context diagrams from paper N9 were incorporated into 4.1.8, and have subsequently been further updated to better reflect the information model terminology and the text of section 4.1.2 to 4.1.7. There are now only two figures -one a data flow and the other a context diagram. The context diagram for Common Services has been removed with the intention of providing additional text on Common Services (section 4.1.1) in the next version.

Item 24: clarify different font sizes in 3.1.9 (SPM)
Status: All data flow item in figures 4-2 and 4-3 are now the same font size.

### 2.10 Information Model (sections 4.2 and 4.3)

Item 1: LR to investigate where algorithms are contained in the model, by September 30**.** (SPM)
Status: OPEN - Currently algorithms are not mentioned explicitly in section 4.2, but the updated data modeling of the access information should enable this to be more easily incorporated in the next version. Needs discussion.

Item 2: LR to clarify either catalogue information or delete it in the RM, by September 30.(SPM)
Status: The US group agreed that it should be taken from Preservation Description Information as a separate category. It is not treated as a special term in this version.

Item 3: typo in first bullet on page 3-14 ‘This give’ should be ‘This gives’.(SPM)
Status: OPEN - this edit was missed.(page 46 - Provenance Information)

Item 4: change in catalogue information ‘extracted’ into ‘derived’ (SPM)
Status: Catalogue Information has been removed

Item 5: Figure 3-3 needs to be updated to make catalogue information optional (SPM)
Status: Catalogue Information has been removed

Item 6: put references to figures numbers into ‘Offpage’ (SPM)
Status: Attempt to do this proved a logistical problem while number of figures and document organization is still subject to change. Will be done for the final version.

Item 7: sub-setting function needs to be included (in storage?) - proposal will be made by the editors (SPM)
Status: Sub-setting has been addressed under Dissemination (Process Data subfunction 4.1.7) and section 4.3.5.

Item 8: change throughout document ‘descriptive records’ into ‘descriptors’ (SPM)
Status: Descriptive Records no longer exists.

Item 9: change term ‘access aids’ (SPM)
Status: We have not found a better term.

Item 10: change term database management into data management (SPM)
Status: Completed

Item 11: type below Fig 3-5 ‘between’ (SPM)
Status: Not clear what was desired here. However, this material on archive holding and associated descriptions has been rewritten in any event.

Item 12: clarify text below Figure 3-6 (SPM)
Status: Completed - This text addressed a view on collection descriptive records. This material has been rewritten and is presented in sections 4.2.3.1.2 and 4.2.3.1.3.

Item 13: change fig. 3-7 to only present up to descriptor level (SPM)
Status: Completed - Figure 4-18.

Item 14: LR to change throughout the document AIP into AIU and vice versa, by July 1. (SPM)
Status: Completed. AIP, which was the smallest ‘unit’ in WB-1, was changed to AIU. However, AIP was retained as the parent class of both AIU and AIC. Thus, SIPs go to AIPs (either as AIUs or AICs) which go to DIPs.

Item 15: Check all OMT figures (2-1, 2-2, 2-3, 2-4, ..) in the document for the strange character string, such as in Figure 2-4 ‘1/97-002’ and delete it in all figures (USB)
Status: Will be done for the final version

Item 16: Revise figure 3-9 as follows: (SPM)

- use OMT methodology
- make figure more readable
- move access box on the side where dissemination box is
- no lines through access - should have defined data flow in and out
- what are the data objects between consumer and access?
- ‘catalogue metadata’ should be changed into ‘descriptors’

Status: OPEN. No updates have been made (now figure 4-19) due to a lack of editorial time and prioritization. We hope to bring a revised draft to the Panel 2 meeting.

Item 17: Scan CNES and BNSC comments on OAIS Version 8 and address all comments in next version, by July 1.
Status: Completed: This has been done, and agreed updates which have not been completed yet were included above.

### 2.11 Migration (section 5)

While no specific edits were recorded, it was clear that the material was preliminary. Substantial updates have been provided, and an effort has been made to relate it to the information model terminology. It needs solid review.

### 2.12 Archive Classifications (section 6)

Item 1: Add text to section 5 to explain why these classifications criteria matter, by September 30.
Status: OPEN - No updates were made due to a lack of editorial resource.

### 2.13 Illustrative Scenario (section 7)

While no specific edits were agreed, it was recognized that the example chosen limited the audience and thus a more broadly based example was incorporated. In addition, an attempt has been made to more fully use the information model terminology. It is recognized that further updates are needed.

### 2.14 Annexes

Item 1: CH/GMP to write a scenario according to the scenario template (US workshop proposed edits), by September 23.
Status: Completed and incorporated into Annex A.

## 3 Initial List of Issues

**Issue 1:**
Currently an AIP is specialized to an AIU or an AIC. Is this sufficiently understandable and useful to keep?

**Issue 2:**
It now looks like the document would flow more smoothly if the data model discussion followed a brief introduction to the functional entities and preceded the detailed functional model discussion, much as has been done in section 2. This would allow an even cleaner use of terms in the detailed functional model section.

**Issue 3:**
It was proposed that ingest does validation/verification and administration does review. However some feel it is more natural to think of all of this activity as a part of Ingest, which is a very extensive and wide ranging activity in many cases. It requires discipline experts to participate in some conversions, for example. See the current write-up on the 'review' subfunction. This is not to suggest that Administration doesn't have the final say, but we think this is not the main activity of interest here.

**Issue 4:**
It is proposed that Archival Storage 'error checking' should include also a checking function for degeneration in the storage. We need discussion on what is seen as practical in this regard.(section 4.1.3)

**Issue 5:**
There was a proposal that a Submission Agreement should be generated with Administration. Currently this is in Ingest. Should this be identified primarily as an Administration function, or should it be in Ingest where there will be the expertise needed to handle most Submission Agreements, apart from formal approvals?

**Issue 6:**
Where should the development of Finding Aids be addressed? There is a Prepare Finding Aids function in Access because, again, we think it is more natural to associate Access with this function than with Administration.

**Issue 7:**
There was a desire to include the concepts of proprietary access and volume control in the Access section. These are not mentioned explicitly now. Do they need to be?

**Issue 8:**
What should be done with Annex C, the matrix?

**Issue 9:**
Currently algorithms are not mentioned explicitly in section 4.2, but the updated data modeling of the access information should enable this to be more easily incorporated in the next version. This needs discussion.

**Issue 10:**
Since the Preservation Description Information also needs to be preserved, where is the Fixity for this information? Is it a part of the Packaging Information (PI)? It seems logical that it would be, in which case the PI may break down into some parts which, under Repackaging, are not to be altered either! Or, does this argue that the PDI should carry the FIXITY for the package as a whole?

---end---